图书情报工作
第66卷 第7期 2022年4月

口述历史档案资源知识图谱与多维知识发现研究*

邓君 王阮
吉林大学商学与管理学院 长春 130012

摘 要：[目的/意义]基于知识图谱的口述历史档案资源知识发现是知识发现在数字人文领域的新尝试，为资源细粒度关联、语义化查询、个性化探索提供新路径。[方法/过程]以南京师范大学抗战老兵口述资料中心数据为数据源构建抗战老兵口述历史档案资源知识图谱，基于图谱实例，从项目整体概况、事件主题关系、社会网络关系、时空网络关系等层面开展多维知识发现研究。[结果/结论]以知识图谱为代表的数字人文技术方法为知识发现研究提供有力的工具支撑，为人文资源深度开发注入了全新动能。

关键词：知识图谱 口述历史档案 知识发现
分类号：G270
DOI：10.13266/j.issn.0252-3116.2022.07.001

1 引言

数字时代，技术的革新发展对人类文化、经济等各方面产生着深远影响。在这一时代背景下，人文资源的知识组织与数字技术的深度交互，催生了知识生产新模式，也推动着学科建设交叉融合。在国家政策导向下，新文科旨在打破传统人文社会学科建设壁垒，建构跨学科、多面向的人文社科新体系。数字人文强调数字技术与人文研究互融共通，借助新技术、新工具实现学科交叉新融合，拓宽人文研究新视野，丰富人文知识新内容。由此可见，新文科与数字人文均强调打破学科分野，实现跨学科、跨领域协同合作。伴随"数字"与"人文"的深度契合，研究者们不再局限于传统的定性分析，而是借助知识图谱等多样可视化方法创新驱动人文研究新范式，加速新文科建设步伐。

当前，学科内部诸要素及属性的演变为数字人文与图书情报与档案管理（以下简称"图情档"）学科的深度交互式研究提供了前提条件[1]。口述历史档案作为一种重要的人文资源，是记录历史记忆的"活化石"，也是中华文化的一种重要传承载体。《"十四五"全国档案事业发展规划》明确指出，鼓励开展口述材料、新媒体信息的采集[2]。面对卷帙浩繁的口述历史档案资源，如何实现资源深度聚合、关联及知识发现亟待研讨。

传统知识发现模式下，口述历史档案资源在网络中呈现分散状态，用户难以将纷繁复杂的资源串联集成，只能通过人工浏览方式对所需资源反复点击，频繁检索，费时耗力，无法实现快速捕获、关联与发现，致使资源价值不能得以有效发挥。数字环境下，以知识图谱为代表的数字人文技术方法能够突破口述历史档案数据固化藩篱，重塑口述历史档案信息堆栈，实现口述历史档案资源关联聚合，完成"数字化—数据化—智慧化"过渡，呈现数据重组—信息关联—知识发现的递进过程及逻辑演变。

因此，本文在新文科建设背景下，以数字人文视角"切入"人文研究，择取知识图谱这一代表性的数字人文知识挖掘方法，以抗日战争为主题背景，引入抗战老兵口述历史档案资源构建知识图谱实例，透过可视化线条深度关联、挖掘潜在知识，以技术赋能人文，探求"数字技术"与"人文研究"的组配可能性，使传统研究方法得以延伸，打破技术与人文资源壁垒，为资源内容特征揭示、语义深度关联及多维知识发现提供全新路径参考。

2 文献回顾

目前知识发现研究主要聚焦基础理论、技术探索、方法应用和研究趋向4个维度。

* 本文系国家社会科学基金项目"数字人文视角下历史档案资源知识聚合与知识发现研究"（项目编号：19BTQ102）研究成果之一。
作者简介：邓君，教授，中国人民大学档案事业发展研究中心研究员，博士，博士生导师；王阮，助理研究员，吉林大学鼎新学者，博士后，通信作者，E-mail：wangruan18@mails.jlu.edu.cn。
收稿日期：2021-05-27 修回日期：2021-09-25 本文起止页码：4-16 本文责任编辑：徐健

理论层面的知识发现围绕概念内涵、过程、步骤的解读。从定义来看，比较认可的知识发现概念是指从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程[3]。从过程步骤来看，马力等[4]将其分解为数据清理、数据集成、数据变换、数据挖掘、模式评估、知识表示 6 个步骤；靳晓恩将过程概括为数据层、工具层、挖掘层、应用层 4 个层次[5]。

在技术层面，以聚类分析、分类分析、机器学习和神经网络为主要方法。在应用领域，学者们多聚焦知识发现平台系统构建、系统功能解析以及具体案例研究，就如何更好地实现领域知识发现进行深入探索。图情领域以曾建勋、孙宇、刘爱琴、陆韡等学者为代表，对领域知识发现系统及功能进行相关研究。宋雪雁等基于数字人文视角，以王世杰日记为例开发名人日记数字化资源，形成人物关系、地域热点、情感倾向等可视化图谱，为内容分析和知识发现提供路径[6]。在档案领域，孙鸣蕾等详细阐述了名人档案知识图谱构建过程，为名人档案开发提供新思路[7]。杨茜茜基于案例剖析与理论溯源，归纳并提出数字人文视野下历史档案资源整理与开发的路径模型及实施方式，使历史档案资源知识发现价值发挥得到进一步深化[8]。

口述历史档案概念自 20 世纪 80 年代引入我国，研究者们就其内涵与外延展开广泛研讨，试图从本质特点、形成过程、法律认可等方面论证其作为档案的合理合法性[9]。伴随档案数字化进程的推进，学者们聚焦国外口述历史档案建设实践经验，以期为我国特色路径发展提供参考借鉴。针对档案资源的知识发现研究，朱令俊以数据驱动为主轴解析档案知识发现的基本程式，从数据层、逻辑层、应用层、表示层构建档案知识发现的内容框架[10]。高晨翔提出了一种档案学视角下的区域政务微博知识发现模型，旨在对具有文件、档案属性的政务信息资源进行知识化开发[11]。H. F. Yu 等实现了基于缓存存档系统(Internet Cache Archive System, ICAS)的归档文件知识发现[12]。M. C. Pattuelli 等讨论了文化遗产链接数据的生成、处理和集成过程，以口述历史档案为链接数据命名实体的主要来源，描述了数据开发过程如何为研究查询和与遗产数据接触提供新途径[13]。

综上所述，在知识发现领域，国内外尚未形成整体合力，尤其在口述历史档案领域拓荒待垦，故而对口述历史档案资源进行关联化的锚定与挖掘成为知识发现新突破口。随着知识发现研究的深入，其本身也在向数字化、智能化发展，这也必将成为口述历史档案资源深层组织、关联、聚合的重要着眼点。同时，数字人文背景下，知识图谱技术方法也能成为口述历史档案资源知识发现的重要动能与激发点。因此，笔者希冀藉由口述历史档案资源知识图谱实例，通过具象化与之关联的人、事、地、时等数据，联结知识网系以发掘隐性知识并赋能知识发现，完成以数据—信息—知识为生长点的动能转换，为资源开发注入新活力，将静态、平面、塔式的信息栈转变为动态、立体、格式的知识网[14]，重塑知识发现研究新范式。

3 口述历史档案资源多维知识发现路径

口述历史档案的资源挖掘，是利用数字技术进行信息提取，因而对于其主题、人物、事件、背景信息的挖掘尤为重要[15]。知识图谱以图结构揭示语义信息，方便用户通过简单的“线上信息”捕获知识、关联知识并发现知识，能为口述历史档案资源知识发现提供可能途径。

本文以南京师范大学抗战老兵口述资料中心数据(包含国民党、八路军、新四军口述项目共计 1 501 个)为实验数据源，采用自顶向下方法构建抗战老兵口述历史档案资源知识图谱，即首先构建模式层；其次在模式层结构的导引下进行数据层组织；最后将处理完毕的数据导入 Neo4j 图数据库进行知识图谱可视化展示，为最终实现口述历史档案资源多维知识发现奠定基础，见图 1。

4 抗战老兵口述历史档案资源知识图谱构建

4.1 抗战老兵口述历史档案资源知识图谱模式层组织

模式层是数据收集的模式框架。本文依据数据源自行抽取所需要素，架构抗战老兵口述历史档案资源知识图谱模式层。

为确保模式层组织的完整性与系统性，笔者基于分类学的观点，从口述历史档案外部特征、内容特征以及形式特征入手，梳理类及层次关系，设置人物(person)、角色(role)、身份(identity)、事件(event)、任职经历(work experience)、时间(time)、地点(place)、军队编制(military establishment)、项目(project)、权限(rights)、设备(device)11 个大类，同时，大类下设相关子类，见图 2。

属性是对类的描述丰富，通常由属性名称、定义域、值域 3 部分构成，属性名称是描述属性的抽象名词，定义域和值域分别为该属性指向的类和数据类型。

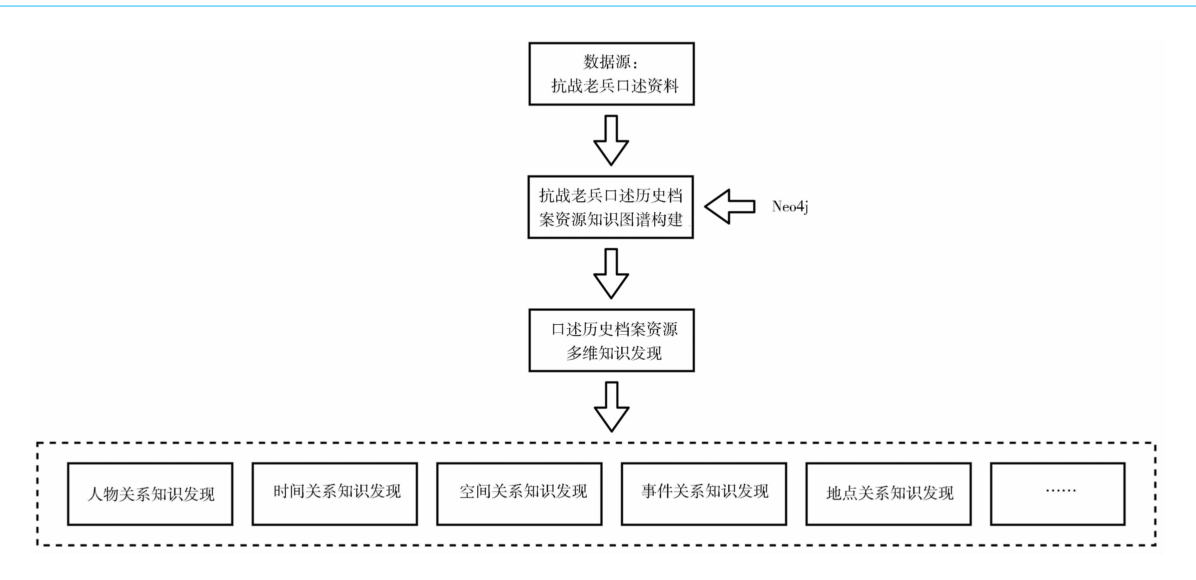


图 1 口述历史档案资源多维知识发现路径——以抗战老兵资料为例

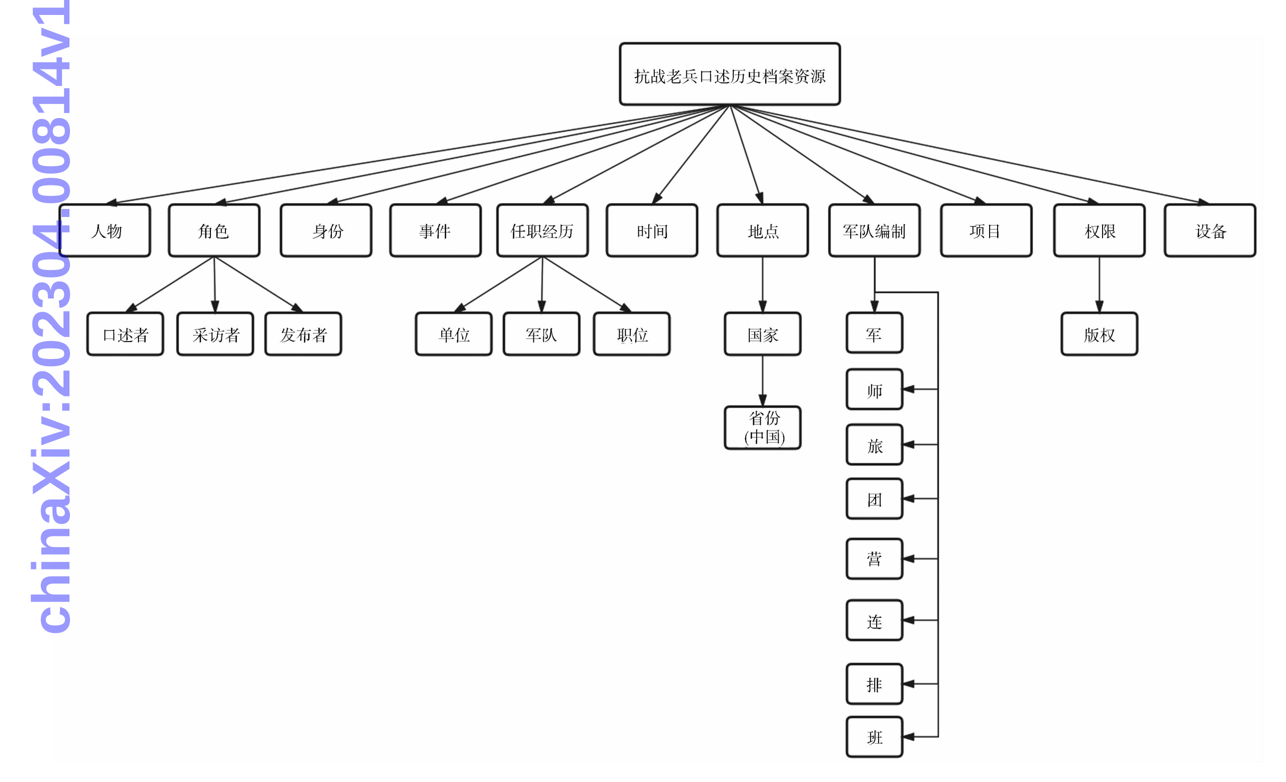


图 2 抗战老兵口述历史档案资源类及层级关系划分

抗战老兵口述历史档案资源属性信息包含 26 个。具体而言，涵盖语言、曾用名、人物姓名、民族、国籍、党派、性别 7 个人物类属性信息；单位曾用名、单位名 2 个单位类属性信息；军职、下设部门 2 个军队属性信息；职位名 1 个职位类属性信息；地名 1 个地点类属性信息；番号(军队)1 个军队编制类属性信息；事件名 1 个事件类属性信息；摘要、录制时长、发布时长、资源大小、资源格式、资源类型、标题、链接、页面底部来源 9 个项目类属性信息；领域内身份认可 1 个身份类属性信息；版权所属 1 个版权类属性信息。

4.2 抗战老兵口述历史档案资源知识图谱数据层组织

抗战老兵口述历史档案资源知识图谱数据层组织包括实体抽取、关系抽取、属性抽取以及知识融合 4 个部分。

4.2.1 实体抽取

实体抽取即通过识别元信息的上下文、字词形式等特征将划分好的字词实体依次分类到既定类别当中。依据实验数据源，抽取符合人物、角色、时间、任职经历等 11 个大类及其下设子类的数据实体，其中设备

类在数据源中并未体现,因此这一大类实体暂无数据填充。

4.2.2 关系抽取

关系抽取即根据实验数据源提取上述类实体间的关系实例,包括关系链接的两个实体及其关系类型信息。以人物类为例,抽取关系类型如表 1 所示:

表 1 关系抽取类型(以人物类为例)

关系类型	定义域	值域
出生时间 date_of_birth	Person	Time
死亡时间 date_of_death	Person	Time
入伍时间 date_of_enlistment	Person	Time
退休时间 date_of_retirement	Person	Time
入党时间 date_of_joining_the_Party	Person	Time
退伍时间 date_of_discharge	Person	Time
人物更名时间 date_of_name_change	Person	Time
籍贯 native_place	Person	Place
现居地 current_residence	Person	Place
曾住地 former_residence	Person	Place
身份 has_identity	Person	Identity
参与 participate_in	Person	Project
角色 has_role	Person	Role
采访 interview	Person	Person
上级 superior	Person	Person
入党介绍人 introducer_to_the_Party	Person	Person
经历 has_experience	Person	Work Experience
亲属 relative	Person	Person
发布 publish	Person	Project

4.2.3 属性抽取

属性有利于实体更加充实、立体。由于同一人物、事件、时间、地点等实体可能存在于不同口述项目,且属性信息也可能存在差异,因此在进行属性抽取时存在属性比对和整合环节。本文在抽取人物属性信息时,仅关注人物 ID 和属性信息的对应情况,即在首次抽取完成后再以人物姓名为筛选项,进行重复人名的信息比对和属性整合工作。

4.2.4 知识融合

经过初步数据筛选及信息抽取,共获得 103 170 条数据,接下来需要对重复信息和冲突信息进行剔除。本实验数据源的知识融合问题,首先体现在数据格式上,即时间格式不统一,例如将“2019 年 09 月 28 日”“1937/7/7”“1940－2－13”等不同日期记载方式统一为“年－月－日”格式。其次表现为数据内容不一致,例如在人名上出现了同人异名的情况,例如“周恩来”“周总理”均指向“周恩来”这一人物实体,这一现象可以通过百科知识图谱对比匹配来实现人名消歧。最后是事件实体存在命名不一现象,即同一历史事件存在多个名称,例如揭开全国抗战序幕的“七七事变”在本次数据源中表述为“卢沟桥事变”“卢沟桥战役”等;而“八一三事变”存在“保卫大上海”“第二次淞沪抗战”“淞沪会战”等名称,因此需要以百科类网站和词典为参考进行事件实体名称统一。

4.3 抗战老兵口述历史档案资源知识图谱可视化

将数据批量导入得到实体节点文件 25 个,关系类文件 64 个,节点数 30 975 个,关系数量 64 037 对。本文选取 Neo4j 图数据库构建抗战老兵口述历史档案资源知识图谱并通过 Browser 工具实现知识图谱可视化。由于界面展示有限,图 3 仅为部分抗战老兵口述历史档案资源数据知识图谱可视化效果。

5 基于知识图谱的口述历史档案资源多维知识发现

5.1 基于项目概况的知识发现

由于口述历史档案资源大多依托项目实践,因此,基于项目概览的知识发现不可或缺。以项目为主线,既可以从宏观视角“勾勒”整体—局部知识图谱,又可以从微观视角“窥探”与之关联的项目—地点知识图谱,为深入展现项目知识发现过程提供路径引导。

为使实体与实体的语义链接信息更加丰富、直观,笔者将 1 501 个口述项目进行知识图谱概览。由此发现,项目网页发布时间多集中于 2016 年、2017 年,项目采集时间跨度较大,从 20 世纪六七十年代至 2015 年,散布于各个年度,且集中于 2015 年,针对大量早年分散的口述历史档案资料进行了抢救性整理并发布,来源丰富,反映史实也更加全面。同时,2018 年后,口述项目基本实现了同年采集同年发布的情况,表明项目整体推进效率增高。项目采集地遍及 20 个省级行政区,涵盖华北地区(北京市、天津市、河北省)、东北地区(辽宁省)、华东地区(上海市、江苏省、浙江省、福建省、江西省)、华中地区(河南省、湖北省、湖南省)、华南地区(广东省、广西壮族自治区)、西南地区(四川省、贵州省、云南省)以及西北地区(甘肃省、陕西省、宁夏回族自治区)。

笔者单独调取“吴岱:沭河两岸军民鱼水情”口述项目,透过知识图谱“连线”表达如下信息:版权所属南京师范大学抗日战争研究中心,项目采集时间为 1987 年,采集地点为北京市,地点所属区域为北京市—

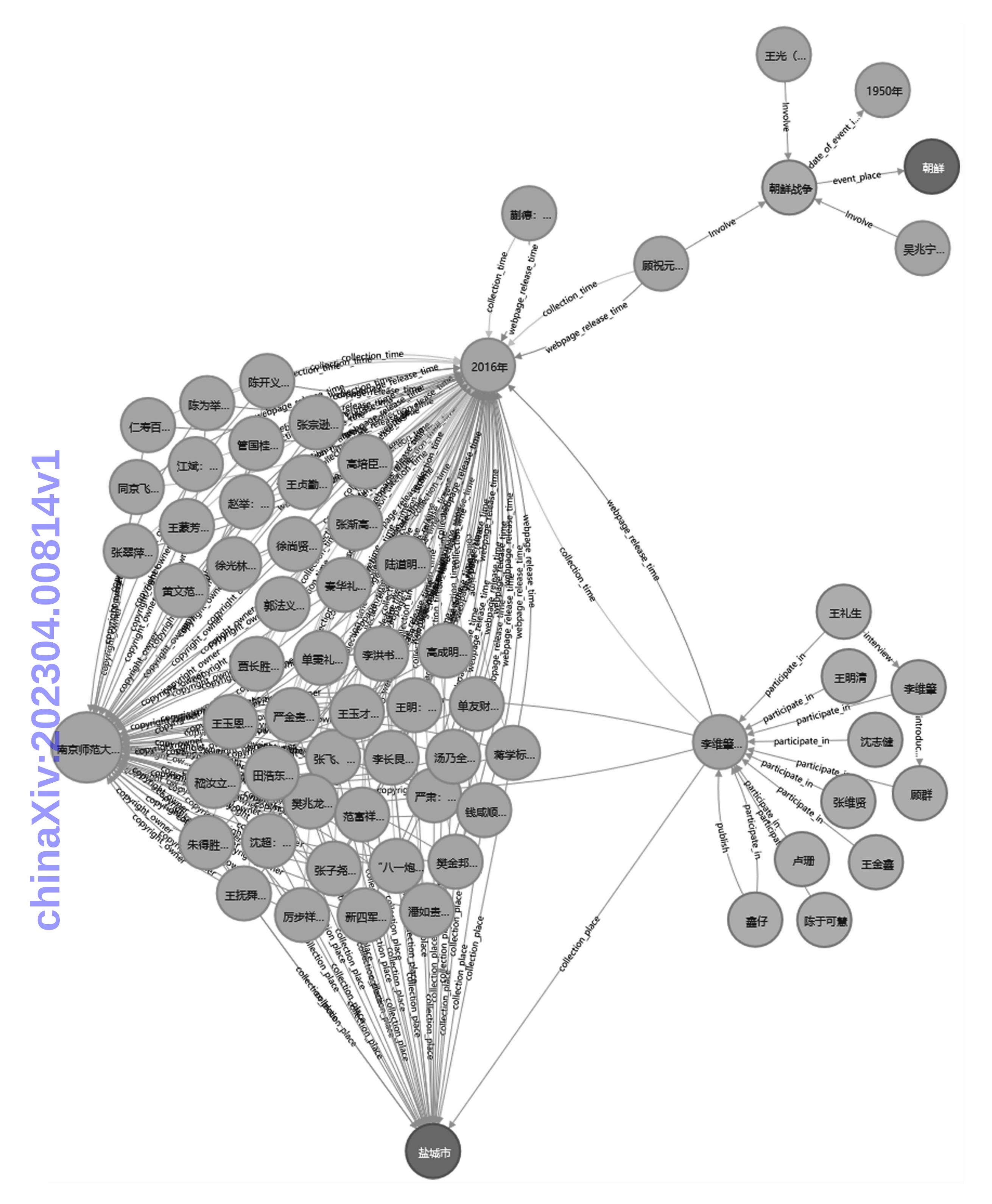


图 3　部分抗战老兵口述历史档案资源数据知识图谱可视化

中国。进一步点击实体，由图 4 展示的属性信息主要是对口述历史档案资源项目外部特征和形式特征的提取总结，如该项目标题为“吴岱：沭河两岸军民鱼水情”，ID 编号为“36175”，URL 链接为 http://lb. njnu. edu. cn/information/262/6477，页面底部来源为“八路军太行纪念馆”，资源大小 21. 22KB，资源格式为 text/html，资源类型为文本、图片。

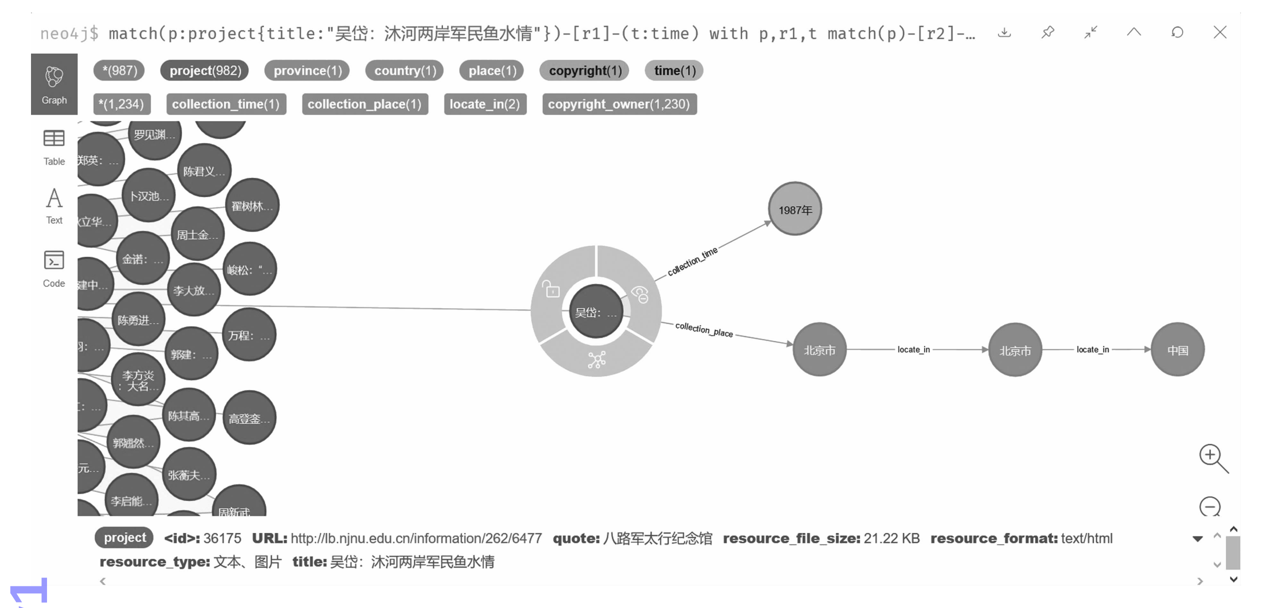


图 4　Neo4j 图数据库“吴岱:沭河两岸军民鱼水情”口述项目操作界面信息

值得注意的是,用户可点击 URL 链接实现从知识图谱可视化端口向口述项目源网页的跳转,这一功能无疑为用户查询检索带来极大裨益。一方面,对项目网页数据的援引是对知识图谱数据真实性、可靠性的进一步检验,用户可参照原文完成对知识图谱的反馈、纠正与更新。另一方面,用户可通过 URL 链接追踪更为翔实的第一人称视角口述信息,为知识链条呈现提供有力资料补充,为多元实体的完整勾画提供保障。

5.2　基于事件主题关系的知识发现

事件作为口述历史档案资源的重要组成部分之一,通常涉及人物、时间、地点等信息,这些辅助元素可以简明扼要地概括事件梗概。由于本文数据源为抗战老兵口述历史档案资源,因此基于事件主题关系的知识发现多为抗战历史事件。为探究事件内在联系,发掘事件潜藏信息,笔者以事件实体为中心,分别对事件—项目关系、事件—时间关系、事件—地点关系进行解构。

5.2.1　事件—项目关系

提取事件—项目关系知识图谱并在图谱下方对相关实例予以细粒度展示,如图 5 所示。从外围区域看,项目—事件知识图谱多呈现一对一关系(即一个口述项目对应一个事件,如口述项目“徐式昌:黄埔抗战老兵的百年沉浮”中只提到了“浙赣战役”1 个事件)和一对多关系(即一个口述项目对应多个事件或一个事件存在于多个口述项目,如口述项目“常端健:第一支枪从敌人手中抢来”中包含了“马义川战役”“堤洞战役”“杏树岗战役”“范马寨战役”4 个事件;不同口述项目如“徐向前:粉碎日军对晋东南的九路围攻”“徐深吉、吴富善:动地军歌唱凯旋——忆响堂铺伏击战”均包含了“神头岭战斗”这一事件),未形成大面积的关系互连,仅散落在图谱外部。

从内部区域看,多线条的交织使得图谱中心出现了大范围多对多关系,存在大面积聚集,某些事件大量重复出现于不同口述项目,或某一口述项目叙述了多个事件,如七七事变、淞沪抗战、车桥战役、台儿庄战役、平型关大捷、湘西会战、衡阳会战、济南战役、太原战役、百团大战、辽沈战役、平津战役、西安事变等,间接表明这些事件对抗战史实研究的重要性,有利于进一步验证或发现相关历史事件,更全面把握当时的时局态势,为研究者进一步深入研究开辟新的问题域、提供新线索[6]。

5.2.2　事件—时间关系

调取事件—时间关系知识图谱,调用函数获取事件—时间分布图(见图 6),发现历史事件在时间上呈现分散和聚合并存的状态,小部分年度仅含 1 个或 2 个历史事件,多为抗战尚未开始或已经结束的年度。其他年度出现了事件聚合,集中在 1937 年－1949 年。说明本文数据源以全面抗战时期的历史事件为主(即发生于 1937 年至 1945 年时间段的历史事件,如七七事变、上海保卫战、临沂之战、长沙第一次会战等),以解放战争期间的历史事件为辅(即 1946 年至 1949 年期间的历史事件,如平津战役、淮海战役、上海战役、杭州战役等)。

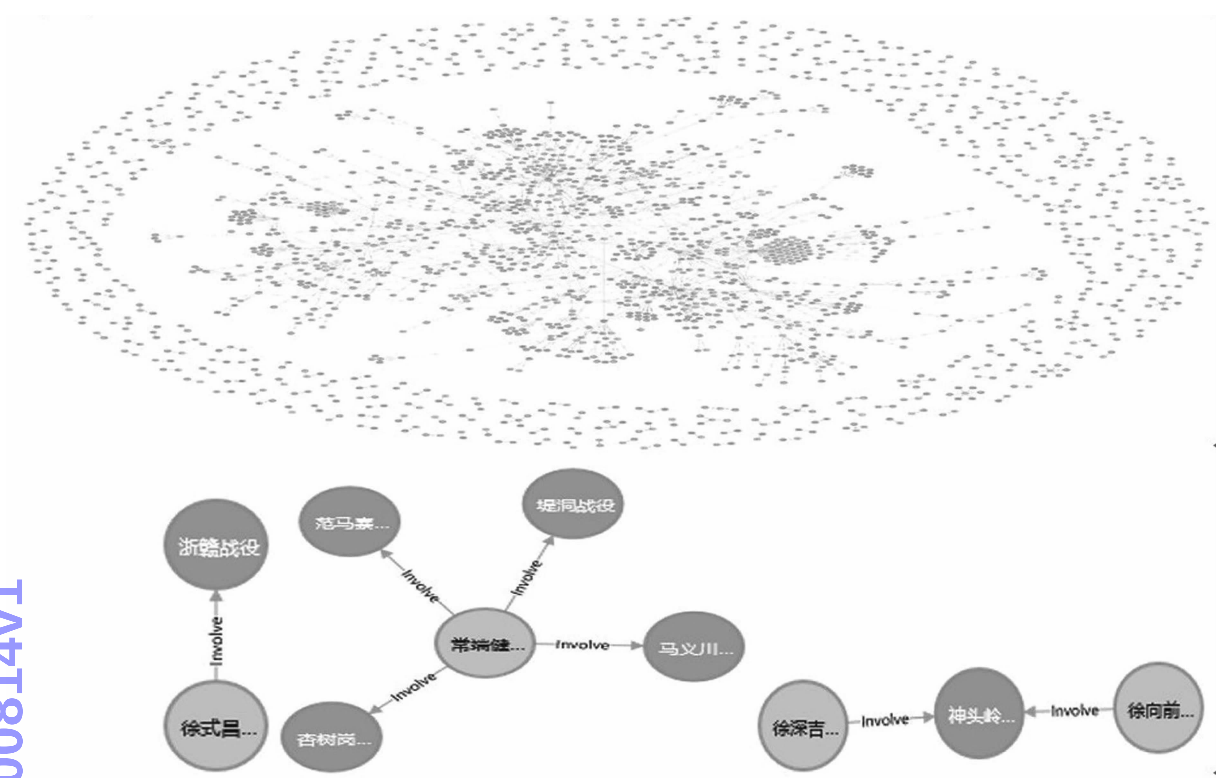


图 5 事件—项目知识图谱及实例展示

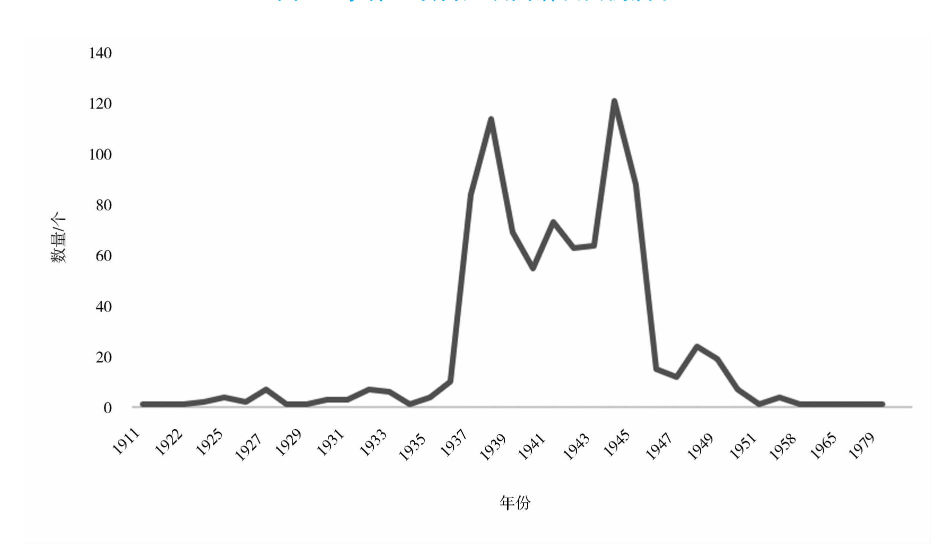


图 6 事件—时间关系分布

5.2.3 事件—地点关系

图 7 为事件—地点关系知识图谱。如图谱中心所示，本文数据源涉及的历史事件集中在山东省、江西省、浙江省、天津市、北京市等地区，通过 count 函数统计，国内省级区划事件数量比例排名前 5 为江苏省（14.4%）、上海市（13.8%）、湖南省（10%）、湖北省（8.8%）、广东省（8.8%）。此外，该知识图谱外围还涉及口述者提及的发生在国外的事件，如朝鲜、印度、缅甸。其中，远征印缅同时发生在印度和缅甸。由此，侧面印证了缅甸战场是中国和太平洋两大抗日主战场的战略结合地带，在一定程度上说明透过知识图谱有利于发现事件—地点知识关联，有助于回溯、审视并深入史学研究，推动人文研究纵深发展。

5.3 基于社会网络关系的知识发现

社会网络是一种基于点和边的揭示社会个体成员之间因社会活动互动而形成的相对稳定的关系体系。基于社会网络关系的知识发现即建立人物—关系映射，形成人物共线关系知识图谱，展现和印证相关人物

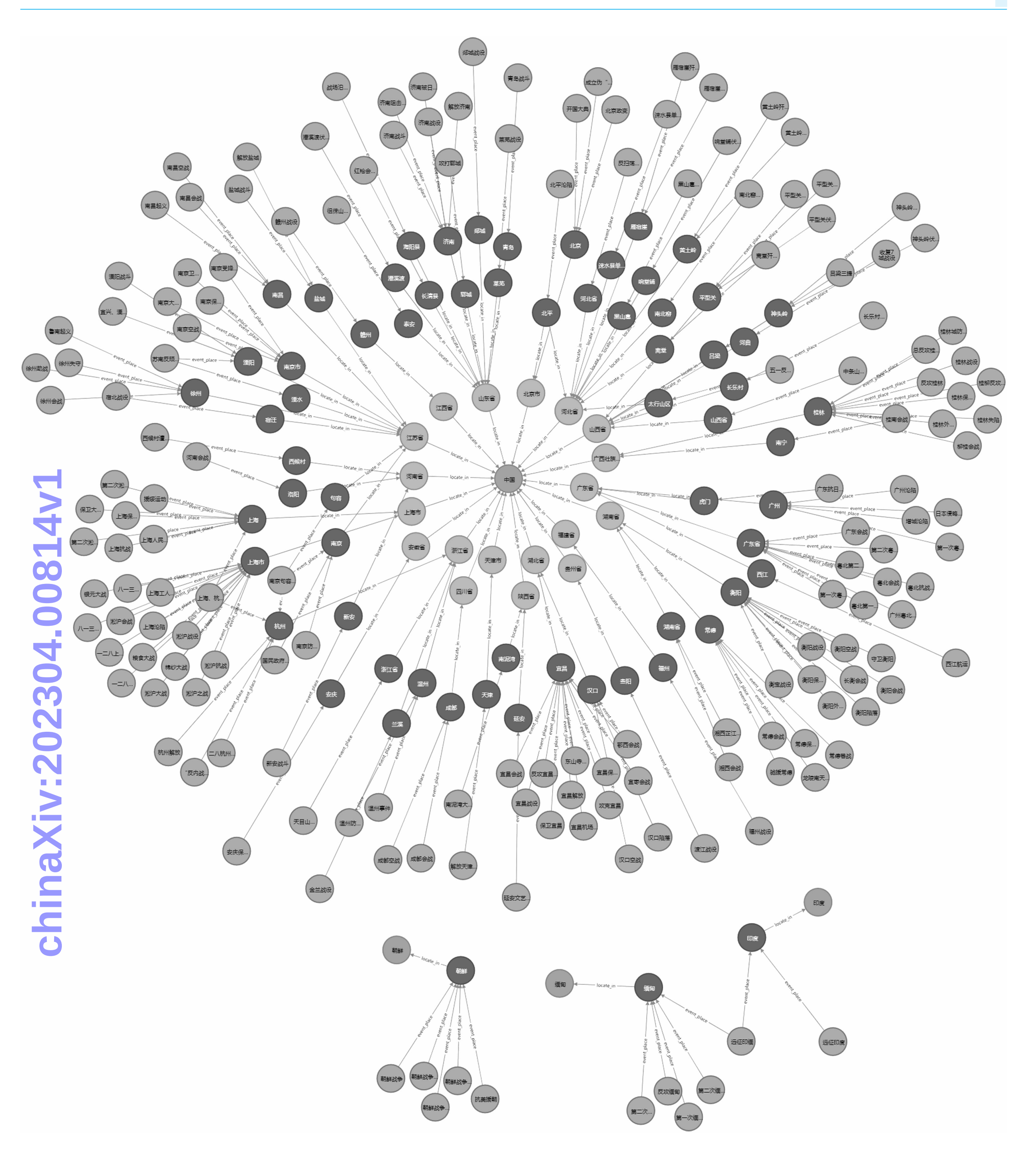


图 7　事件—地点关系分布

在不同历史时期的社会关系状态和重要性。本文包含上下级关系的权力群体分析以及亲属关系的人物梳理两个层面。

5.3.1　上下级关系的权力群体分析

上下级关系是口述项目社会网络中的重要组成部分，本文采用中介中心度(centrality)来衡量个体在社会网络中的核心作用(代码见图 8)。当关系图谱某一实体中介中心度较高时，意味着该实体对于其他人物的把控力较强，进一步说明其权力、地位、身份处于核心领导地位。将人物实体中介中心度值排序，抽取排名前 50 的人物，见表 2。同时，按党派将人物划分为两大类别，一类为中国共产党核心人物，另一类为国民党核心人物。

表 2　人物中介中心度排名

排名	person_name	score	排名	person_name	score	排名	person_name	score	排名	person_name	score
1	朱德	309.0	14	张云逸	39.5	27	皮定均	24.0	40	张牧	18.0
2	卫立煌	234.0	15	陈明韶	37.0	28	李子芳	24.0	41	彭雪枫	17.5
3	陈诚	185.0	16	郑洞国	33.0	29	邓子恢	24.0	42	熊绶春	15.0
4	聂荣臻	126.0	17	谭震林	32.5	30	杨勇	22.0	43	梁植基	15.0
5	欧致富	100.0	18	邓龙光	32.0	31	李先念	22.0	44	张爱萍	15.0
6	杨成武	84.0	19	阎锡山	30.0	32	文安庆	21.0	45	夏光	15.0
7	刘伯承	75.0	20	杜聿明	30.0	33	罗炳辉	21.0	46	李弥	14.0
8	彭康	64.0	21	宋任穷	30.0	34	何克希	20.5	47	王凌云	13.0
9	舒同	60.0	22	汪达之	26.0	35	桂永清	20.0	48	康生	13.0
10	解蕴山	46.0	23	孙立人	25.0	36	张发奎	20.0	49	叶剑英	12.0
11	彭德怀	45.0	24	左权	25.0	37	徐向前	18.0	50	朱理治	12.0
12	粟裕	45.0	25	彭德怀	25.0	38	毛泽东	18.0			
13	邓小平	42.0	26	罗卓英	24.0	39	杨得志	18.0			

```
CALL gds. betweenness. stream({nodeProjection: 'person',
relationshipProjection: 'superior'})
YIELD nodeId, score
RETURN gds. util.asNode (nodeId). person_name
AS person_name, score
ORDER BY score DESC
```

图 8　人物中介中心度计算实现代码

在中国共产党核心人物中，一部分为第一代中央领导集体成员，例如朱德、邓小平、彭德怀、毛泽东等；另一部分为军事方面的核心人物，如聂荣臻、刘伯承、彭德怀、粟裕、李先念、彭雪枫等。在中国国民党核心人物中，以陈诚、卫立煌、阎锡山、张发奎等为代表，如图 9 所示。因国民党内部存在着众多派系，派系的发展斗争也是政治权力更替转移的结果，因而，对抗战时期中国国民党核心人物的知识发现研究分析可从派系着手。本文数据源主要涉及黄埔系和地方系两大派系。黄埔系包含杜聿明、罗卓英、薛岳等重要人物，该图谱展示了与杜聿明关联的邱清泉、陈朋、丁占福、罗卓英等人物网络关系。地方系涉及晋军系和粤军系两派系。其中，晋军系以阎锡山为中心，李梦源、赵承绶、蔡荣寿、荣鸿胪、商震为核心人物。粤军系如张发奎派系，以黄福荫、左洪涛、梁植基、李磊夫为主要人物。此外，笔者发现与汪精卫关联的人物有其下级细俊六、陈公博、李士群、陈璧君、柴山、周佛海，叔侄关系的汪屺，且汪精卫与陈璧君不仅存在夫妻关系还存在上下级关系。摒弃了以往单调的文字叙述，知识图谱的多阶线条为人物关系快速捕获及完整勾勒提供了导引，这是知识图谱进阶知识发现的魅力所在。

5.3.2　亲属关系的核心人物梳理

亲属关系为社会生活形态下的基本社会关系构成单位[16]，基于亲属关系的社会网络支配着整个社会结构及社会秩序。为确保亲属关系划分和称谓的准确性和科学性，本文择录较为完整的《亲属称呼辞典》[17]作为参考资料，对亲属关系进行识别、更正、整合。同时，考虑到各国亲属体系和亲属称呼各异，特别是我国亲属关系划分之细、界定之多，因此，除普通亲属关系外，为便于统计分析，本文将数据源提及的同宗关系、同学关系、同事关系、师生关系、旧相识等关系一同纳入亲属关系范畴，具体见表 3。

在明确亲属关系类型和对应名称的前提下，本文以中国人民解放军创始人之一的叶挺为例，人物亲属关系图谱见图 10。总体而言，以叶挺为核心起点，包含与之关联的人物及人物间的亲属关系共计 4 种。可获得包括其妻子李秀文、儿女叶华明、叶正明、叶正大等直接语义信息以及由多级人物关系递增产生的隐藏人物关联，如任光及其妻子徐韧均为叶挺的下级。以叶挺的下级李子芳为例，李子芳与施荷塘存在母女关系，与李菲螺为父子关系。由此可见，知识图谱将叶挺—任光—徐韧—李子芳等人物实现了“串联”，为人物关系深层揭示与发现提供优化路径，即实体“连线”可直击人物关系，实现多阶知识发现。

5.4　基于时空网络关系的知识发现

时空网络关系包括时间和空间两个维度。时间与空间的结合勾画了一个多维立体的关联世界，人物、事件等平面实体通过时间和空间的加成，实现了由单维

邓君，王阮. 口述历史档案资源知识图谱与多维知识发现研究[J]. 图书情报工作,2022,66(7):4－16.

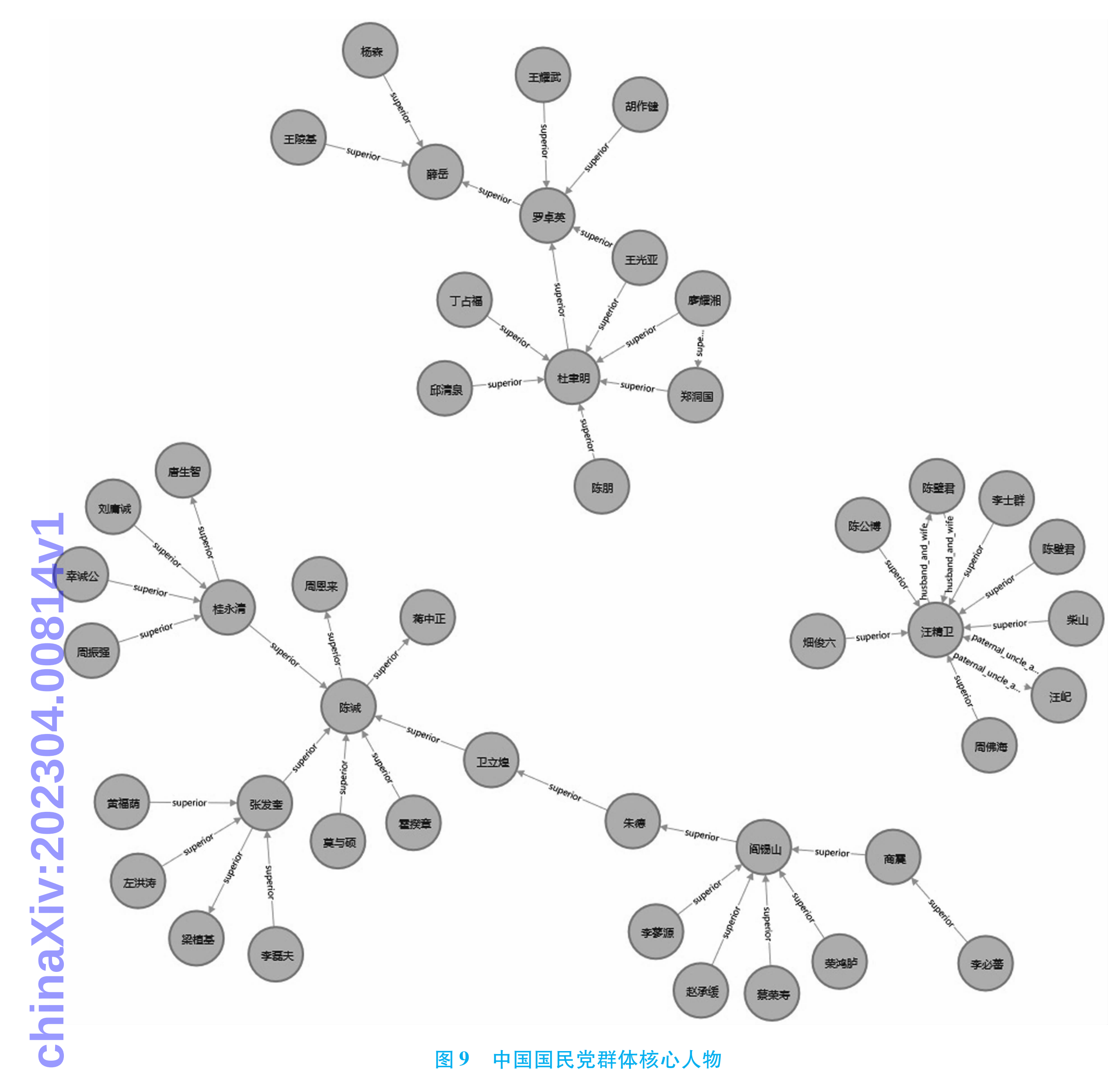


图 9　中国国民党群体核心人物

表 3　亲属关系中英文对照表

中文关系	英文对照	中文关系	英文对照
父女	father and daughter	前夫妻	ex-husband and ex-wife
父子	father and son	公媳	father-in-law and daughter-in-law
母女	mother and daughter	翁婿	father-in-law and son-in-law
母子	mother and son	叔嫂	brother-in-law and sister-in-law
兄弟	sibling	祖孙	grandparent and grandchild
姐妹	sister	祖侄	granduncle and grandnephew
兄妹	brother and sister	舅甥	maternal uncle and nephew
姐弟	sister and brother	姑侄	amitate
表兄弟	brother from mother’s side	叔侄	paternal uncle and nephew
表兄妹	brother and sister from mother’s side	堂叔侄	cousin-nephew relationship
堂兄弟	brother from father’s side	亲戚	relative
堂姐弟	sister and brother from father’s side	同宗	clansman
堂兄妹	brother and sister from father’s side	同学	classmate
族兄弟	the family brothers	同事	colleague
情侣	couple	师生	teacher and student
夫妻	husband and wife	旧相识	acquaintance
未婚夫妻	fiancée and fiancé		

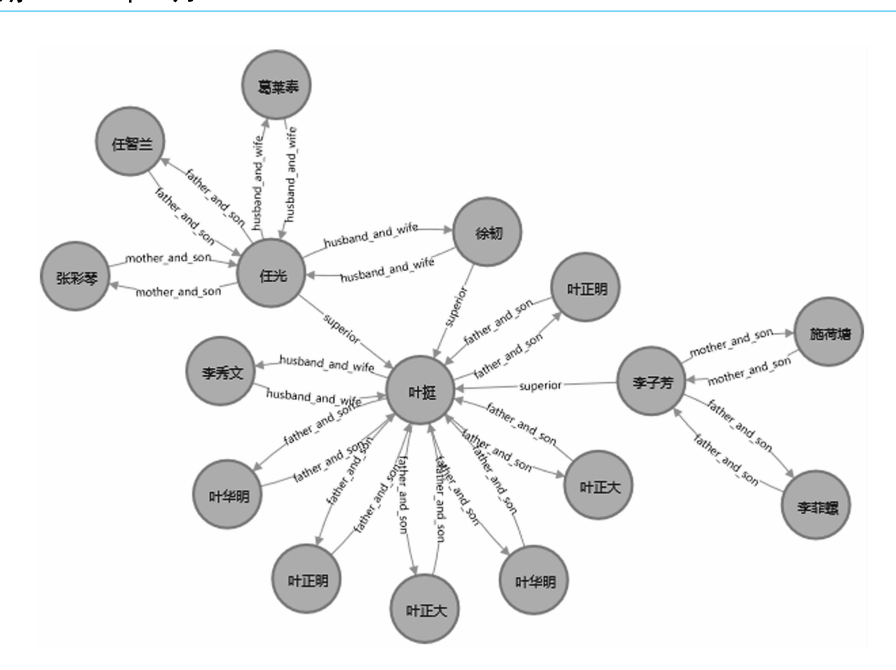


图 10　以“叶挺”为例的人物亲属关系图谱

向多维的实境实现，为探寻人物、事件等实体随时间、空间维度变化的迁移路径与规律提供了解决方法。

5.4.1　基于社会关系的人物空间分布分析

社会关系网络的空间分布特征以人物为核心实体。探究人物空间关系分布，一方面可以聚焦某一地域，获取人物聚集区域，从而从侧面反映一个地区参与抗战的人物数量多少；另一方面可以聚焦人物群体特征，探寻群体社会关系网络空间分布规律。将人物籍贯地数据映射至柱状图（见图 11）可实现人物—空间信息“聚类”。由此发现，人物社会关系空间分布较为分散，大多集中在华东地区，以江苏省和浙江省为两个主要聚集地，同时在华中、华南、西南地区也有少量分布。

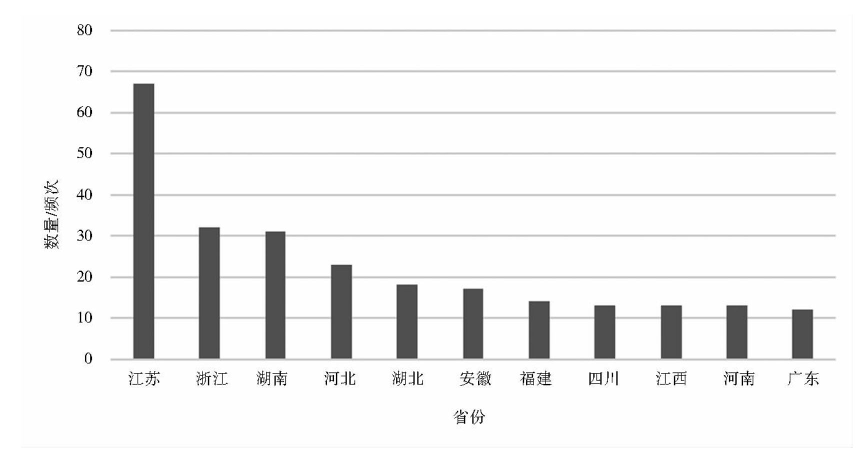


图 11　人物籍贯地分布柱状图

5.4.2　基于任职经历的人物时空迁移轨迹分析

人物的任职经历变化同步产生时空迁移，有关迁移轨迹的知识发现同样折射出抗战时期的社会变迁特征，对人文研究具有重要参考价值。本文人物任职经历包括单位和军队两部分，存在任职调动以及转业关系，并且这些关系伴随着时间和空间特征，构成了人物时空任职经历体系。

以单个人物任职经历为例，笔者调取开国上将之一吕正操将军的任职经历，通过人物时空迁移轨迹实现关联信息知识发现，见图 12。

可以发现，吕正操有两段任职经历（军队和单位）。最早在辽宁省国民党东北军第 16 旅任训练教导队队长，并于 1933 年离职转入国民党东北军第 16 旅第 647 团任团长，而后又成为国民党东北军第 116 师第 691 团团长。1937 年，吕正操从国民党东北军离职，同年加入共产党，成为人民自卫军司令员，任职地点也

邓君，王阮. 口述历史档案资源知识图谱与多维知识发现研究[J]. 图书情报工作,2022,66(7):4－16.

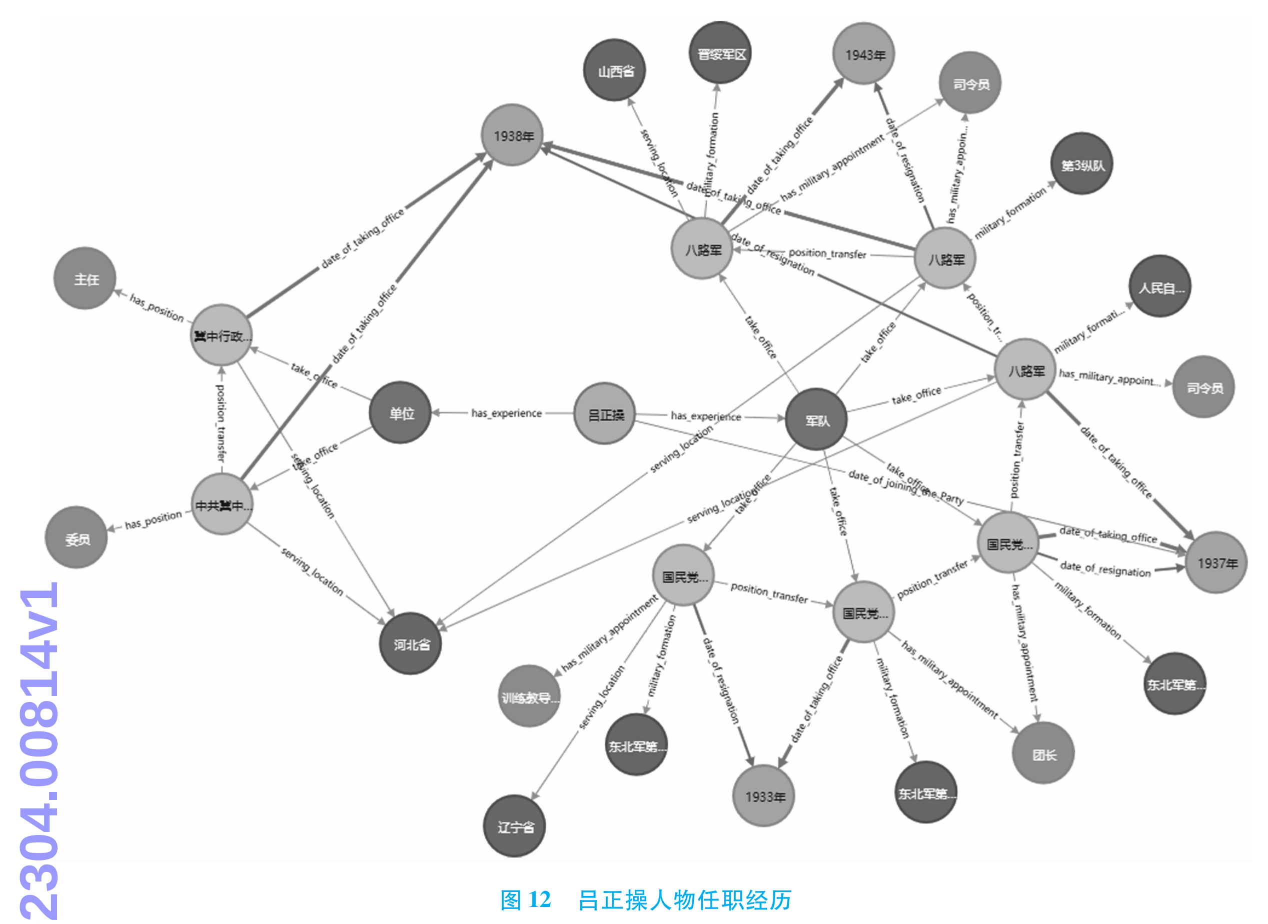


图 12　吕正操人物任职经历

从辽宁省迁移至河北省。至 1938 年，吕正操离开人民自卫军，成为八路军第 3 纵队司令员，并接连任职中共晋中区党委委员，冀中行政主任、公署主任。1943 年，吕正操调任至晋绥军区任军区司令员，任职地点迁移至山西省。

6　结语

本文以抗战老兵口述历史档案资源为实验数据源构建知识图谱，形成了与口述历史档案数据实例密切相关的高维知识表征空间，可以辅助研究者直接“定位”所需信息，佐证了以知识图谱为代表的数字人文技术方法在知识发现研究中的实用价值与应用价值，完成了口述历史档案资源数据级、知识级的存储及知识关联，由此形成的口述历史档案资源知识图谱关联网系可作为后续多维知识发现的重要“原材料”。可以说，知识图谱以其强大的语义处理能力和开放组织能力，为知识化组织和智能化应用奠定了基础，基于知识图谱的多维知识发现研究为挖掘人文资源潜藏内容提供了新途径，丰富并拓展了知识发现研究体系。笔者希冀以点带面，以局部联动整体，为人文资源知识发现创新突破注入新思维、新视角、新动能。同时，探索跨类型、跨渠道的多维知识发现体系建设[18]，助力图情档资源兼容并包，互联互通。

由于界面截图展示有限，本文知识图谱的完整性和操作性欠佳。因此，在未来研究中，笔者将试图找寻知识图谱与网页前端的结合点，为用户提供在线检索操作及更多的互联功能展示，拓宽人文研究路径，推动人文资源在理论创新、方法变革、模式探索等方面的层层递进，助力人文研究向多开发主体参与、个性化检索、深层次挖掘、交互式融通的新模式迈进。

参考文献：

[1] 牛力，高晨翔，张宇锋，等. 发现、重构与故事化：数字人文视角下档案研究的路径与方法[J]. 中国图书馆学报，2021，47(1)：88－107.
[2] 国家档案局. “十四五”全国档案事业发展规划[EB/OL]. [2021－11－08]. https://www.saac.gov.cn/daj/yaow/202106/899650c1b1ec4c0e9ad3c2ca7310eca4.shtml.
[3] FAYYAD U M. Advances in knowledge discovery and data mining[M]. Cambridge: MIT Press，1996.
[4] 马力，焦李成. 一种基于粗集理论的知识发现系统的研究与设计[J]. 微电子学与计算机，2003(3)：8－12.
[5] 靳晓恩. 基于知识发现的数字图书馆用户信息知识化管理研究[J]. 图书馆学研究，2013(17)：18－20，34.
[6] 宋雪雁，崔浩男，梁颖，等. 数字人文视角下名人日记资源知识

发现研究——以王世杰日记为例[J]. 情报理论与实践,2021,44(6):105-111.

[7] 孙鸣蕾,房小可,陈忻. 数字人文视角下名人档案知识图谱构建研究——以作家档案为例[J]. 山西档案,2020,(6):79-88.

[8] 杨茜茜. 数字人文视野下的历史档案资源整理与开发路径探析——兼论档案管理中的历史主义与逻辑主义思想[J]. 档案学通讯,2019(2):17-22.

[9] 潘玉民,叶徐峥. 论口述历史档案是档案的理由[J]. 北京档案,2016(5): 14-17.

[10] 朱令俊. 数据驱动下档案知识发现的路径研究[J]. 档案与建设,2020(2):30-34,13.

[11] 高晨翔. 档案学视角下区域政务微博的知识发现模型研究[D]. 西安:西北大学,2019.

[12] YU H F, CHEN Y M, TSENG L M. Archive knowledge discovery by proxy cache[J]. Internet research: electronic networking applications and policy, 2004, 14(1):34-47.

[13] PATTUELLI M C, HWANG K, MILLER M. Accidental discovery, intentional inquiry: leveraging linked data to uncover the women of jazz[J]. Digital scholarship in the humanities, 2017, 32(4):918-924.

[14] 潘威. "数字人文"背景下历史地理信息化的应对——走进历史地理信息化 2.0 时代[J]. 云南大学学报:社会科学版,2018,17(6):80-87.

[15] 高淞,王向女. 数字人文视域下口述历史档案资源开发利用研究[J]. 山西档案,2021(3):61-70.

[16] 莘欣. 从婚姻法到民法典——中国特色亲属制度研究[A]//上海市法学会.《上海法学研究》集刊(2020 年第 9 卷总第 33 卷)——民法典婚姻家庭妇女权益保护文集[C]. 上海:上海市法学会,2020:5.

[17] 鲍海涛,王安节. 亲属称呼辞典[M]. 长春:吉林教育出版社,1988.

[18] 阮光册,夏磊,周萌葳. 跨媒体智能视角下的知识服务探析[J]. 情报理论与实践,2021,44(7):79-85.

作者贡献说明:

邓君:论文修改;

王阮:提出论文整体研究思路与框架;数据收集与分析;论文撰写与修改。

Research on Knowledge Map and Multidimensional Knowledge Discovery of Oral History Archives Resources

Deng Jun Wang Ruan

School of Business and Management, Jilin University, Changchun 130012

Abstract: [Purpose/Significance] Knowledge discovery of oral history archives resources based on knowledge map is a new attempt of knowledge discovery in the field of digital humanities, which provides a new path for fine-grained association, semantic query and personalized exploration of resources. [Method/Process] Taking the data of the Anti-Japanese War Veterans Oral Data Center of Nanjing Normal University as the data source, the knowledge map of oral history archives resources of anti-Japanese war veterans was constructed. Based on the map examples, multidimensional knowledge discovery research was carried out from the aspects of overall project overview, event theme relationship, social network relationship, space-time network relationship and so on. [Result/Conclusion] Digital human technology represented by knowledge map provides powerful tool support for knowledge discovery research and injects new kinetic energy into the deep development of human resources.

Keywords: knowledge map oral history archives knowledge discovery